Group 41
Phase 4

The code contains 4 functions for preprocessing of the data:

- def transform_data()
- def normalize_data()
- def encode_data()
- def feature_selection()

The functions transform_data, normalize_data, and encode_data take the original dataset as an argument. The function feature_selection takes the encoded data as an argument.

def transform_data():

- This function checks if each column of the data frame is empty and takes an appropriate action.
- Numerical values like the year, death_number, age_mortality, death_percentage, and death_rank are filled with the median of those values if there are empty fields.
- Empty values in country and state are filled with NA, denoting not available.
- The fields age_range, sex, death_description, and mortality_code are filled with the general values ‘Age at time of death, all ages’, ‘Both sexes’, ‘Other causes of death’, ‘[Other]’ respectively.
- Returns the transformed data.

def normalize_data():

- This function normalizes the numerical data from the dataset.
- The attributes year, death_number, age_mortality, death_percentage, and death_rank are normalized.
- Normalization is done using the MinMaxScaler() method from Sklearn.
- Returns the normalized data.

def encode_data():

- This function encodes the categorical data from the dataset.
- The attributes country, state, age_range, death_description, and mortality_code are encoded.
- This function uses the OneHotEncoder() method from Sklearn to one-hot encode the data.
- Returns the encoded data.

def feature_selection()

- This function determines the feature selection for the encoded data.
- This function uses the VarianceThreshold() method from Sklearn.
- The variance threshold was selected to be 0.05. Any higher variance thresholds would only cause 1 or no attributes to be selected.